Smallset Timelines with smallsets: CHEAT SHEET

smallsets is a tool for visually documenting and communicating data preprocessing decisions. It builds a Smallset Timeline figure [3] based on preprocessing code in an R, R Markdown, or Python script. Users must first add structured comments, with building instructions, to the preprocessing code.

Structured comments

Structure

# smallsets instruction data caption(text) caption

Mandatory instructions

start start code tracking and take first snapshot
end end code tracking and take final snapshot

Optional instructions

snap take intermediate snapshot after next line
resume add resume marker after next line

Main functions

Smallset_Timeline(data, code, ...) builds a Smallset Timeline

sets_sizing() for adjusting sizing parameters, including column names, caption text, snapshot data, and legend items

sets_spacing() for adjusting spacing parameters, including caption space, column name rotation, and number of figure rows

sets_labelling() for adjusting the colours of the column names and snapshot data

Demo dataset and code

The smallsets package comes with example data and preprocessing code, which are used to illustrate how the package works, such as in the next section.

Synthetic dataset

s_data 100 observations and eight variables (C1-C8)

Preprocessing scripts

s_data_preprocess.R basic preprocessing scenario in R
s_data_preprocess.Rmd basic preprocessing scenario in R Markdown
s_data_preprocess.py basic preprocessing scenario in Python
s_data_preprocess_4.R includes additional snapshot
s_data_preprocess_resume.R includes resume marker

Steps to build a Smallset Timeline

The demo dataset s_data and preprocessing code s_data_preprocess.R are used to illustrate the process.

Step 1

Add structured comments to the preprocessing code in your R, R Markdown, or Python script, specifying snapshot points and captions.

File: s_data_preprocess.R

```r
define smallsets

# script start
# code do something
# code do something
# code do something
# code do something
# script end
```

Step 2

Run the main smallsets command to build a Smallset Timeline for your dataset and preprocessing code.

Smallset_Timeline(data=s_data, code="s_data_preprocess.R")

Smallset selection

To select the small number of rows from the original dataset used in the visualisation, you can use one of three selection methods available in the Smallset_Timeline() command.

rowCount number ofSmallset rows (5-15)
rowSelect Smallset row selection method

= 1 → coverage model (Gurobi required)
= 2 → coverage + variety model (Gurobi required)
= NULL → random sampling

Warning. This method has long runtimes for large datasets. See [2] for workarounds.

References

[1] CRAN reference manual
cran.r-project.org/web/packages/smallsets/smallsets.pdf

lydia.lucchesi.github.io/smallsets/articles/smallsets.html

paper providing a detailed discussion of Smallset Timelines, the Smallset selection optimisation problems, and two case studies with example Smallset Timelines
doi.org/10.1145/3531146.3533175

PrintedData = TRUE show the data values in Smallset snapshots
GhostData = TRUE plot blank rows/columns after data deletions
MissingDataTints = TRUE use colour tints to highlight missing data

Smallset_Timeline(data=s_data, code="s_data_preprocess.R")

Data has not changed
Data has been edited
Data has been added
Data will be deleted

colour blind safe
B&W printer friendly

To customise the information in a Smallset Timeline and its appearance, you can set different parameters in the Smallset_Timeline() command. See [1] for the complete list of parameters.

```r
colours = c(1, 2, 3)
```